GutUZH at CLEF2025 BioASQ Task 6: a Method of SOTA Performance with the Best Results at GutBrainIE NER Subtask 1

Notebook for the GutUZH Lab at CLEF 2025

Jinyi Han[1,*,†], Yeyang Liu[2,†]

[1]University of Zurich (UZH), Department of Computational Linguistics, Andreasstrasse 15, 8050 Zurich, Switzerland
[2]University of Zurich (UZH), Department of Informatics, Binzmühlestrasse 14, CH-8050 Zürich, Switzerland

Abstract

This paper presents the GutUZH group's participation in CLEF2025 BioASQ Task 6 "GutBrainIE Subtask 6.1 - Named Entity Recognition". We addressed two main questions: 1) the ability of a pre-trained PubMed biomedical NER model to adapt to the more specific subdomain of the gut-brain axis, given limited annotated data, 2) the performance upper bound of our BiomedBERT+CRF model when combined with various training strategies. Our team achieved the best results among 17 participating groups, not only in the official evaluation measure (Micro-F1) but also in Macro-F1 and Micro-Precision, with 4.8% improvement over the official Micro-F1 baseline provided by the task organizers. Our findings demonstrate that domain-specific pre-trained BiomedBERT models remain strong competitors for medical NER tasks. Additionally, data augmentation and ensemble methods proved effective in enhancing our model's performance.

Keywords

GutBrainIE CLEF 2025, Biomedical NLP, gut-brain interplay, BioASQ, BiomedBERT, BNER, CEUR-WS

1. Introduction

The exponential growth of biomedical literature necessitates advanced Natural Language Processing (NLP) techniques to extract and structure valuable information. Named Entity Recognition (NER) serves here as a foundational task, identifying and categorizing key entities within unstructured text. This paper details the participation of team "GutUZH" in the GutBrainIE CLEF 2025 shared task, specifically Subtask 6.1: Named Entity Recognition. The GutBrainIE challenge [1, 2], Task 6 of the BioASQ CLEF Lab 2025[3], is situated within the context of the EU-supported HEREDITARY project, focusing on extracting structured information from biomedical abstracts concerning the gut microbiota and its relations with Parkinson's disease and mental health.

The GutBrainIE shared task aims to foster the development of Information Extraction (IE) systems capable of assisting experts in understanding the complex gut-brain interplay. Subtask 6.1, Named Entity Recognition, requires participants to identify and classify text spans (entity mentions) within PubMed abstracts into one of 13 predefined categories. These categories include entities such as bacteria, chemical, microbiota, disease, disorder, or finding (DDF), anatomical location, and others crucial to the gut-brain domain. The expected output format for each identified entity is a tuple: (entityLabel;

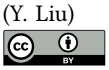
CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding authors.

†These authors contributed equally.

† Han compared LLM fine tuning against suitable pre-trained language models, boosted the model performance by changing input representations, modifying the model's output architecture, inference methodologies, provided modular and extensible codebase, drafted the initial manuscript with plots.
Liu enhanced model performance through optimized fine-tuning strategies on learning rate, data quality, data format, and ensemble-based data augmentation

jinyi.han@uzh.ch (J. Han); yeyang.liu@uzh.ch (Y. Liu)
https://github.com/VirginiaPoe/GutBrainIE_2025_PubMedBERTcrf (J. Han); https://github.com/yyLeaves/gutbrainie25 (Y. Liu)



entityLocation (title/abstract); startOffset; endOffset).This structured output facilitates downstream applications like relation extraction and knowledge graph construction.

NER plays a critical role in extracting information from biomedical literature. Effective BiomedNER systems can accelerate research by highlighting key concepts and their mentions in scientific texts. Biomedical NER (BNER) however, presents significant challenges. These include the inherent complexity of biomedical terminology, which often involves long, multi-word expressions, specific chemical names, and variations in nomenclature. Entity ambiguity, where terms can have different meanings depending on context, is another hurdle. Furthermore, creating a high-quality annotated datasets for training NER models is labor-intensive and expensive. The GutBrainIE task itself introduces this data scarcity challenge through its dataset structure, which comprises of varying annotation quality: Gold (expert-annotated), Platinum (expert-validated), Silver (student-annotated and weakly curated), and Bronze (model generated and no manual revision) [1].

- **Our Contribution: An Iterative Approach to State-of-the-Art Performance**
  This paper describes the system developed by team "GutUZH" for Subtask 6.1 of GutBrainIE CLEF 2025. The core contribution lies in the systematic and iterative development process undertaken to achieve optimal performance. Our journey involved several stages:
  1. Fine-tuning BiomedBERT *microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext* [4], a language model pre-trained on biomedical literature, using the combined Gold, Platinum, Silver datasets provided by the organizers.
  2. Subsequent architectural enhancements with significant performance boost from dev Micro F1 at 0.5781 to test Micro f1 at 0.8120, surpassing the official baseline. Enhancements includes inserting title/abstract indexing as a novel feature during data pre-processing, adding dropout layer, linear classifier and Conditional Random Field (CRF) layer to improve sequence labeling.
  3. Continuous performance improvement to reach test Micro F1 at 0.8408, through domain-specific data augmentation, random seed ensembling, and data quality control.
- **Structure of the Paper**
  The paper is organized as follows: Section 2 shortly reviews related work in biomedical NER, focusing on transformer-based models, CRF layers, training techniques. Section 3 provides a detailed description of the proposed system, including each stage of its development. Section 4 outlines the experimental setup, including dataset characteristics, evaluation metrics, and implementation details. Section 5 presents the experimental results, including performance comparisons. Section 6 discusses the findings and considers limitations.

2. Related Work

Transformer-based Models for BNER Transformer architecture [5] revolutionized the field of NLP. Models based on Transformers, such as BERT (Bidirectional Encoder Representations from Transformers), have achieved state-of-the-art performance on a wide range of NLP tasks, including NER. The key innovation of Transformers is the self-attention mechanism, which allows the model to weigh the importance of different words in a sequence when representing a particular word, thereby capturing long-range dependencies effectively.

For biomedical applications, domain-specific versions of BERT have been developed by pre-training on large biomedical corpora. BioBERT [6] and BiomedBERT[4] are examples. BiomedBERT, in particular, is pre-trained from scratch on PubMed abstracts and full-text articles from PubMed Central (PMC). It is highly attuned to the language and terminology of biomedical literature. As of 2025, Transformer-based encoder models, especially those derived from BERT, remain the backbone of state-of-the-art Named Entity Recognition (NER) systems. While large language models (LLMs) like GPT-4 [7], LLaMA [8] are being explored, they have not yet displaced transformer-based encoder models fine-tuned on specific NER tasks as the most reliable or computationally efficient solution—especially in low-resource or

high-accuracy settings. [9, 7, 10, 11, 8]. This context informed our decision to prioritize from an LLM-based approach to fine-tuning BERT model.

Conditional Random Fields (CRF) in NER Conditional Random Fields (CRFs) are a class of statistical modeling methods often applied to structured prediction tasks, including sequence labeling problems like NER. While Transformer models like BERT excel at generating powerful contextual token representations, their standard output layer (e.g., a softmax over labels for each token) makes independent classification decisions for each token. This can lead to label inconsistencies, such as predicting an "I-Disease" (Inside-Disease) tag without a preceding "B-Disease" (Begin-Disease) tag, which violates common tagging schemes like BIO (Begin, Inside, Outside) or BIOES (Begin, Inside, Outside, End, Single). A CRF layer added on top of a neural network encoder (such as a BiLSTM or a Transformer) addresses this limitation by considering the dependencies between adjacent labels in the sequence [12, 13]. The CRF layer learns transition scores between pairs of labels and combines these with the emission scores (output from the encoder for each token) to find the globally optimal sequence of tags for a given input sentence [14]. This global normalization helps ensure that the predicted tag sequences are valid and coherent. The combination of a Transformer encoder with a CRF layer (Transformer-CRF) has become a widely adopted and effective architecture for NER tasks [15, 16]. The rationale is clear: Transformers provide rich contextual features, while CRFs enforce sequential constraints and improve the structural integrity of the output predictions.

Data Augmentation in BNER Data scarcity is a persistent challenge in developing robust NER systems. BNER high-quality annotation is costly and time-consuming [17]. Traditional augmentation methods include synonym replacement, back-translation, noising techniques, and sampling methods. Many generic augmentation techniques can inadvertently alter the meaning of the text or, more critically, invalidate the existing entity labels (e.g., deleting part of a named entity or replacing a word within an entity with a non-entity word) [18]. Therefore, domain-aware or task-specific augmentation strategies are often preferred in BNER [19, 20]. As explored in this work, related unlabeled or pseudo-labeled texts from the same domain was incorporated.

3. Methodology

Our system for Named Entity Recognition (NER) is focused on fine-tuning a pretrained biomedical Transformer model, microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext [4]. The input features for the model are PubMed article titles and abstracts. In our main configuration, these segments are concatenated with special tokens [TITLE] and [ABSTRACT], separated by a single space, and then tokenized using the Transformer's vocabulary. Entity spans with the text are annotated using the standard BIO (Beginning, Inside, Outside) tagging. The tokenized input is processed through BiomedBERT's embedding layer and its multi-layer Transformer encoder to generate contextualized representations for each token. These token representations are fed into a dense linear layer, which projects them into the NER tag space. Emission scores are produced for each possible BIO (Beginning, Inside, Outside) tag. To model dependencies between adjacent tags and ensure globally optimal tag sequences correctly, a CRF layer is applied on top of these emission scores. During inference, the final sequence of BIO tags for a given input is determined by the CRF layer using the Viterbi decoding.

3.1. Iterative Model Development

We adopted an error-driven iterative approach for our NER architecture.

- **Model Comparison**
  Initially, we considered BioClinicalBERT [21], BioBERT [22], DistilBERT [23], DeBERTa-v3-large [24], and BiomedBERT (previously named PubmedBERT) [4]. Our final choice of BiomedBERT was guided by the hypothesis that a model deeply immersed in the target domain's language would yield superior performance for fine-grained entity recognition. Specifically, the rationales are as follows.

- **Domain-Specific Pretraining:** BiomedBERT's pretraining on a massive corpus of biomedical literature provides it with a rich understanding of biomedical terminology, syntax, and semantic relationships [4]. The base model is pre-trained on 14 million PubMed abstracts with 3 billion words (21 GB), fulltext model is increased to 16.8 billion words (107 GB). This is crucial for bio-clinical NER tasks where accurately identifying specialized entities (e.g., diseases, genes, chemicals) is challenging. Models pretrained on general corpora, like DistilBERT or DeBERTa-v3-large, often require more extensive fine-tuning to adapt to the nuances of a specialized domain like biomedicine and may still not capture domain-specific knowledge as effectively [6, 25, 26].
- **Vocabulary Similarity:** The vocabulary learned by BiomedBERT during pretraining is aligned with the biomedical domain. We hypothesize that common and rare gut-brain interaction terms are more likely to be well-represented within its embedding space. General-domain models whose vocabularies might tokenize biomedical terms as out-of-vocabulary instances [27].
- **Advantage over other Domain-Specific Models:** While BioBERT and BioClinicalBERT are also pretrained on biomedical or clinical text, BiomedBERT (specifically the PubMedBERT lineage) was chosen because of its reported superior performance in several benchmark biomedical NLP tasks, including NER, as documented in existing literature [28, 29, 30]. Although BioClinicalBERT excels in clinical contexts, our task focuses on biomedical research literature rather than clinical notes directly, we hypothesized that BiomedBERT's PubMed pretraining potentially is more advantageous.

- **Baseline Configuration and Initial Challenges**
  Taking BiomedBERT as the base model, the intuitive system configuration was directly concatenating article titles and abstracts (separated by a single space) as input. This text was tokenized, processed by the encoder, the token embeddings were fed into a linear classification layer followed by a softmax function for independent BIO tag prediction, with cross-entropy loss as the loss function. However, the dev inference output showed challenges, with dev Micro-F1 at only 0.5781, as shown in Table 1, where Experiment is named BiomedBERT.

  - **Semantic Misclassifications**
    We observed semantic misclassifications. Terms with contextual proximity or conceptual overlap to target entities were incorrectly labeled. Words like "potential" were wrongly labeled "microbiome". The model had a tendency of overgeneralization, it incorrectly assigned specific entity labels to broader conceptual terms, such as labeling "Infectious agents" as DDF (Disease or Disorder of Function) or "genetic susceptibility" as gene. These terms were semantically similar, but cannot meet the entity span's criteria.
  - **Localization Errors**
    There was also localization errors. Entity spans that appeared both in the abstract and title was confusing to locate.
  - **Incomplete Entity Spans**
    Another challenge observed in the baseline model's output was the prediction of incomplete or fragmented entity spans. For instance, phrases such as "characterization of" were sometimes incorrectly identified as entities, despite not representing complete, standalone concepts according to the GutBrainIE annotation guidelines. This indicated that the model had difficulties with precise boundary detection, potentially starting an entity based on strong local signals (e.g., the word "characterization") but failing to identify the correct endpoint.
  - **BIO Sequence Constraint Violations**
    The model produced invalid BIO tag sequences that violate the fundamental constraints of the BIO tagging scheme during training. For example, an "I-microbiome" tag appeared immediately after "B-food".

It could be that the base model treats each token's label prediction as an independent classification problem, without considering the sequential dependencies and constraints inherent in the BIO tagging scheme. Each token is classified without awareness of the previous token's predicted label. The model lacks explicit mechanisms to enforce BIO tagging rules. Although the cross-entropy loss optimizes individual token predictions, it doesn't enforce global sequence consistency across the entire tag sequence.

- **Structural Improvements**
  - **Structural Markers for Localization Improvement**
    As localization errors was observed, our first architectural refinement was to insert explicit structural markers. We added '[TITLE]' to title segments and '[ABSTRACT]' to abstract segments before their concatenation and tokenization. These markers were integrated as new tokens into the tokenizer's vocabulary. The hypothesis was that these distinct markers would provide the model with clearer contextual cues about which section of text it processed.
    No dramatic location identification improvement was observed after adding these special tokens, showned in Table 1 BiomedBERT + spetok.
  - **Enhancing Span and Semantic Coherence with a CRF Layer**
    We hypothesized that incomplete/unreasonable span errors and semantic problems stemmed from the token-level predictions being made independently by the softmax classifier, without considering the global coherence of the tag sequence across the entire input.
    Our next key architectural enhancement was to integrate a CRF layer. This layer was positioned on top of the linear classification layer, which was modified to produce emission scores (logits) for each BIO tag per token, rather than direct probabilities via softmax. The CRF layer is then trained to learn valid transition probabilities between adjacent tags (e.g., enforcing that a 'B-DDF' tag is likely followed by an 'I-DDF' or an 'O' tag, but not directly by a 'B-human' tag if they represent distinct entity types). It finds the most likely sequence of tags for the entire input, not just the most likely tag for each isolated token.
    During inference, the Viterbi algorithm (VD) is employed by the CRF to decode the globally optimal sequence of BIO tags for the entire input, considering both the emission scores from the linear layer and the learned tag transition probabilities.
- **Performance of the Refined System Configuration**
  This iteratively developed model, with BiomedBERT as the base model, the '[TITLE]' and '[ABSTRACT]' as special markers, a linear classification head, and the CRF layer with Viterbi decoding (VD), achieved a Micro-F1 score of 0.8117 on our development set, shown in Table 1 BiomedBERT + spetok + CRF + VD. This performance surpassed the official baseline for the first time. Together with the training configurations, our model underwent subsequent different training strategies. Our model architecture is illustrated in Figure 1.

3.2. Training Strategy

Our training incorporated several key strategies:

- Fine tuning BiomedBERT model pretrained on medical corpus: By fine tuning on a BERT model already familiar with medical text [4], we saved the efforts of injecting medical knowledge into the model by ourselves.
- Differential Learning Rates: We implemented different learning rates for different components of the model:
  - BiomedBERT: 2e-5 (smaller learning rate to prevent catastrophic forgetting)
  - Linear classification layer: 2e-4
  - CRF layer: 2e-3 (higher learning rate to overcome cold start challenges)

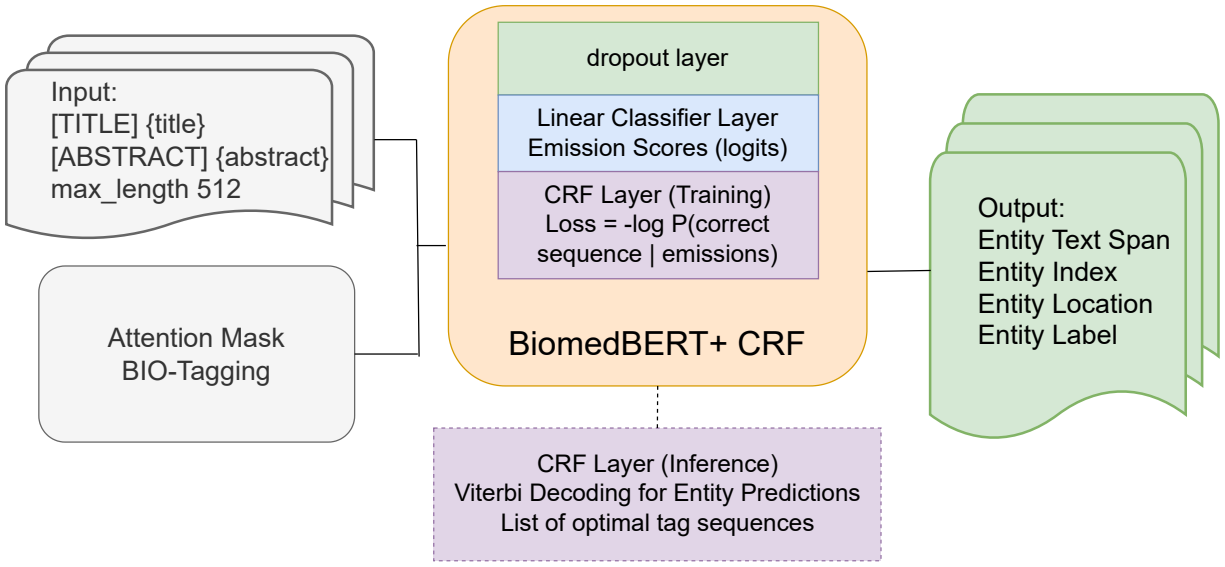


Figure 1: Model Architecture of our refined NER system

This differential approach acknowledges the varying learning dynamics of each component and optimizes their convergence rates accordingly.

- Data Quality Control Experiments: We carefully experiment with Data Quality by
  - Initially training on high-quality labeled data (Platinum, Gold, Silver)
  - Later incorporate Bronze-level data to expand the training corpus
  - Implementing pseudo-labeling for data augmentation
  - Updating Bronze data labels with higher-quality predictions from our improved models
- Model Ensemble of Different Seeds: To enhance robustness and generalization, we implemented an ensemble approach:
  - Training multiple models with different random seeds (11, 17, 42)
  - Averaging the emission and transition matrices across these models
  - Combining their predictions to make more reliable sequence labeling decisions

  This ensemble method consistently improved both macro and micro F1 scores across our experiments.
- Data Augmentation: We experimented enhance of training data through following setups:
  - Pseudo-labeling: We scraped 500 additional PubMed articles on gut-brain axis topics and generated labels using our best-performing ensemble model. These pseudo-labeled examples were then incorporated into the training set.
  - Bronze Data Label Improvement: We updated the labels of Bronze-quality data using predictions from our improved models, effectively upgrading their quality.
  - Continual Training: After initial training on the full dataset, we performed additional training focused exclusively on high-quality data (Platinum, Gold, Silver) to reinforce the patterns learned from the most reliable examples.

Inference Optimization During inference: For abstracts whose tokens exceeds max sequence length, we split it by last period token before length limit and then conduct inference separately to avoid truncation-related performance degradation.

Development Workflow Our development workflow followed an iterative experimental approach:

1. Compare and choose an optimal base model
2. Inference error analysis
3. Structurally improve the base model, by adding dropout layer, linear classifier, CRF layer, while continuously comparing inference errors
4. Exploring different loss functions
5. Systematically vary individual components and parameters
6. Evaluate changes on validation data
7. Integrate successful modifications into the main pipeline
8. Repeat the process with new hypotheses

This methodical approach allowed us to progressively improve our system's performance while maintaining a clear understanding of how each modification contributed to overall results.

4. Experimental Setup

Our NER system is implemented using PyTorch, with pre-trained BiomedBERT weights from the Hugging Face Transformers library. The complete code implementation, including data preparation, training, and inference scripts, is available in our public repository at https://github.com/yyLeaves/gutbrainie25.

Hardware Configuration

All experiments were conducted on the following hardware:

- GPU: NVIDIA GeForce RTX4090 (16GB)
- CPU: Intel Core i7-13900H
- RAM: 32GB DDR4

Training Data Composition

As described in the CLEF 2025 task overview website [1], we used the dataset provided from the biomedical literature focused on research on the gut-brain axis. The data for this task is a set of titles and abstracts of biomedical articles retrieved from PubMed, focusing on the gut-brain interplay and its implications in neurological and mental health. The dataset was organized into different quality tiers:

- Platinum: Expert-annotated and external-reviewed with highest quality annotations
- Gold: Expert-curated high-quality annotations
- Silver: Trained-student-annotated Mid-quality annotations under expert supervision
- Bronze: Automatically labeled with fine-tuned GLiNER as base models

Evaluation Metrics

The task used micro-average F1-score as primary metric for ranking metrics on the final leaderboard, chosen for its ability to handle class imbalances effectively [1]. Other measures such as Macro-P, Macro-R, Macro-F1, Micro-P, Micro-R, Micro-F1 are also provided.

Data Preprocessing

We designed our preprocessing pipeline to handle titles and abstracts separately, as our experiments indicated that this would be beneficial. Here are the key steps:

1. Title/Abstract Tokenization with Special Tokens: We precess the title and abstract part of an article separately as two independent entries, appending the [TITLE] / [ABSTRACT] token to indicate the location of the text.
2. Entity Label Alignment: Character-based entity spans are mapped to BIO labels, handling subword tokenization effects.
3. Input Formatting:

- Initial Format: [CLS]{title tokens} {abstract tokens}[PAD]
- Improved Format 1: [CLS][TITLE]{title tokens} [ABSTRACT]{abstract tokens}[PAD]
- Improved format 2: [CLS][TITLE]{title tokens} [CLS][ABSTRACT]{abstract tokens}[PAD]

4. Padding and Truncation: Sequences are padded/truncated to a fixed length (MAX_SEQ_LEN), initially set as 512, with -100 masking ignored tokens.
5. Label Encoding: Converting BIO tag sequences into numeric labels for model training
6. Multi-Tier Data Integration: Platinum/Gold/Silver/Bronze datasets are concatenated for controlled quality experiments.

Hyperparameter Configuration

We employed the following hyperparameters across our experiments:

- Batch size: 8 during training, 16 during evaluation
- Maximum sequence length: 512 tokens
- Epochs:
  for model development, 20 with early stopping, patience = 3
  for continuous improvement via training stategies, 40 with early stopping, patience = 5
- Optimizer: AdamW
- Learning rates:
  BiomedBERT: 2e-5
  Classification layer: 2e-4
  CRF layer: 2e-3
- Weight decay: 0.01
- Warm-up steps: 500, 0.01
- Loss function:
  For the BERT base model, we employed cross-entropy loss, defined as:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \tag{1}$$

  where N is the number of samples, C is the number of classes, $y_{i,c}$ is the true label (1 if sample i belongs to class c, 0 otherwise), and $\hat{y}_{i,c}$ is the predicted probability for sample i belonging to class c.
  For our optimized models, we utilized CRF negative log-likelihood loss:

$$\mathcal{L}_{CRF} = -\log P(y|x) = -\log \frac{\exp(\text{score}(x,y))}{\sum_{y'}\exp(\text{score}(x,y'))} \tag{2}$$

  where $P(y|x)$ is the conditional probability of the true label sequence y given input x, $\text{score}(x,y)$ represents the compatibility score between input x and label sequence y, and the denominator sums over all possible label sequences y' to ensure normalization.

Experimental Variations

We conducted a series of systematic experiments to optimize our approach:

- Base model BiomedBERT: Training on high-quality data only (Platinum, Gold, Silver), by directly concatenating title and abstract (separated by a single space)
- Unified text format: [CLS][title]...[abstract]
- Structural model improvements on the output: dropout, linear classification, and CRF
- Random seeds (17, 42, 123, 777) and learning rate (2e-5, 3e-5, 1e-5) adjustments to find the optimal model performance
- Uniform learning rate (2e-5) for all model components
- Learning Rate Optimization: Differentiated learning rates for model components

- Multiple random seeds (11, 17, 42) to evaluate stability: Evaluation across multiple seeds to ensure consistent improvement
- Incorporation of Bronze-quality data
- Update Input Format: Separate processing of title and abstract, reduction of information loss due to truncation
- Model Ensemble: Combination of models trained with different random seeds by Averaging of emission and transition matrices
- Addition of 500 PubMed articles with pseudo-labels generated by our best-performing ensemble model
- Bronze Data Quality Improvement: Relabeling Bronze data with our best models: Integration of updated labels into training
- Continual Training: Additional training on high-quality data after initial convergence

5. Results

Our team (GutUZH) achieved the first place on the leaderboard, demonstrating the effectiveness of our systematic approach to biomedical named entity recognition. Our best-performing model achieved a Micro-F1 score of 0.8408 on the official test set, establishing a new benchmark for this task.

The performance metrics of our final system are:

- Macro-P: 0.7950
- Macro-R: 0.7736
- Macro-F1: 0.7613
- Micro-P: 0.8384
- Micro-R: 0.8432
- Micro-F1: 0.8408 (primary evaluation metric)

These results represent a substantial improvement over our base model, which was initially with a Micro-F1 of 0.5781, and eventually at 0.8408, demonstrating an absolute improvement of 26.27%, 4.81% above the official baseline. An overview of all the experimental results is shown in table 1.

Table 1
Experimental F1 Results Summary

Experiment	Seed 42	Seed 11	Seed 17	Ensemble	Test Result
BiomedBERT	Macro: 0.4869 Micro: **0.5781**	Macro: 0.4868 Micro: **0.566**	Macro: 0.4963 Micro: **0.5719**	–	–
BiomedBERT + spetok	Macro: 0.473 Micro: **0.5502**	Macro: 0.4258 Micro: **0.5097**	Macro: 0.4852 Micro: **0.5714**	–	–
BiomedBERT + CRF + VD	Macro: 0.7136 Micro: **0.8042**	Macro: 0.7143 Micro: **0.8108**	Macro: 0.7337 Micro: **0.8028**	–	–
BiomedBERT + spetok + CRF + VD	Macro: 0.7246 Micro: **0.8074**	Macro: 0.7648 Micro: **0.8096**	Macro: 0.7286 Micro: **0.8117**	– –	Macro: 0.7100 Micro: **0.8120**
BiomedBERT + spetok + CRF + VD + LR	Macro: 0.7136 Micro: **0.8087**	Macro: 0.7365 Micro: **0.8156**	Macro: 0.7589 Micro: **0.8316**	–	–
BiomedBERT + spetok + CRF + VD + Bronze	Macro: 0.7139 Micro: **0.8169**	Macro: 0.7513 Micro: **0.8347**	Macro: 0.7522 Micro: **0.8321**	–	–
BiomedBERT + spetok + CRF + VD + Training Format	Macro: 0.7595 Micro: **0.8288**	Macro: 0.7705 Micro: **0.8373**	Macro: 0.7671 Micro: **0.8334**	Macro: 0.7773 Micro: **0.8394**	Macro: 0.7634 Micro: **0.8328**
BiomedBERT + spetok + CRF + VD + Augmentation	Macro: 0.7569 Micro: **0.8379**	Macro: 0.7498 Micro: **0.8400**	Macro: 0.7724 Micro: **0.8408**	Macro: 0.7662 Micro: **0.8457**	Macro: 0.7613 Micro: **0.8408**
BiomedBERT + spetok + CRF + VD + Update Label	Macro: 0.7405 Micro: **0.8374**	Macro: 0.7610 Micro: **0.8310**	Macro: 0.7469 Micro: **0.8279**	Macro: 0.7613 Micro: **0.8414**	–
BiomedBERT + spetok + CRF + VD + Continual Learning	Macro: 0.7598 Micro: **0.8372**	Macro: 0.7367 Micro: **0.8319**	Macro: 0.7495 Micro: **0.8305**	Macro: 0.7667 Micro: **0.8417**	Macro: 0.7686 Micro: **0.8361**

Our systematic experimental approach yielded the following key improvements:

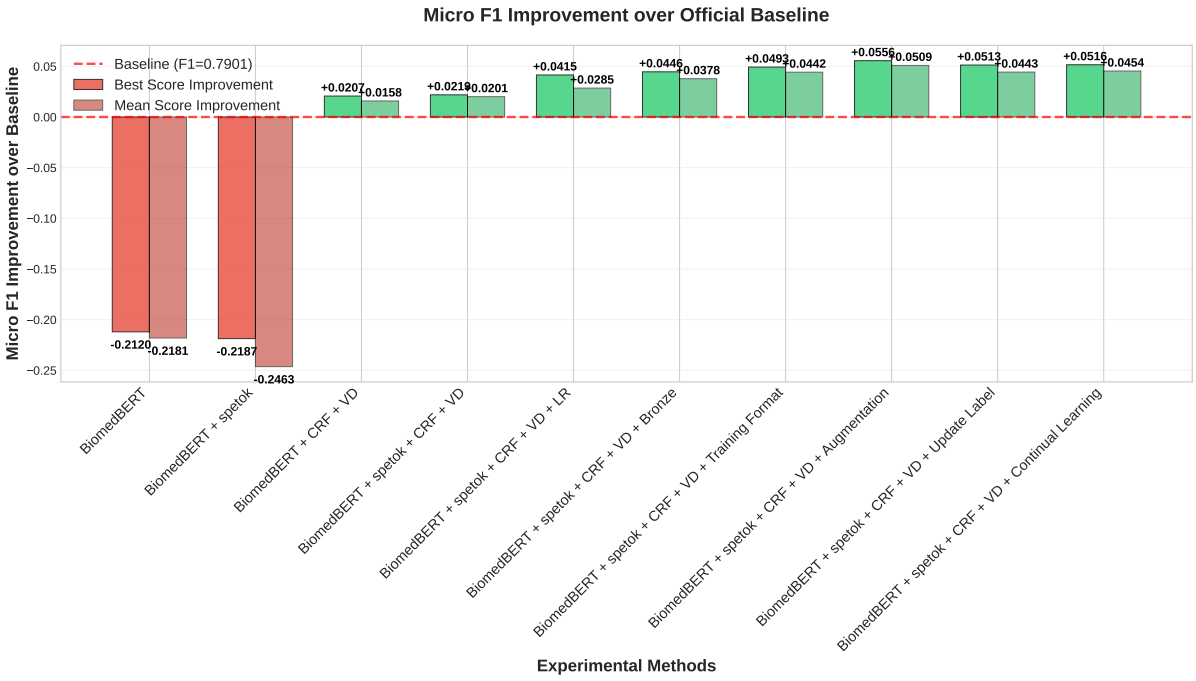


Figure 2: Iterative model development improvements. BiomedBERT serves as the base model, with sequential additions of special tokens (spetok), CRF layer, Viterbi decoding (VD), learning rate optimization (LR), ensemble methods, and training strategies. Final configuration achieves +0.0454 improvement over baseline. Best score (the bar on the left) represents the highest observed micro F1 among all seeds and settings. Mean score (the bar on the right) averages all available micro F1 values, reflecting method stability.

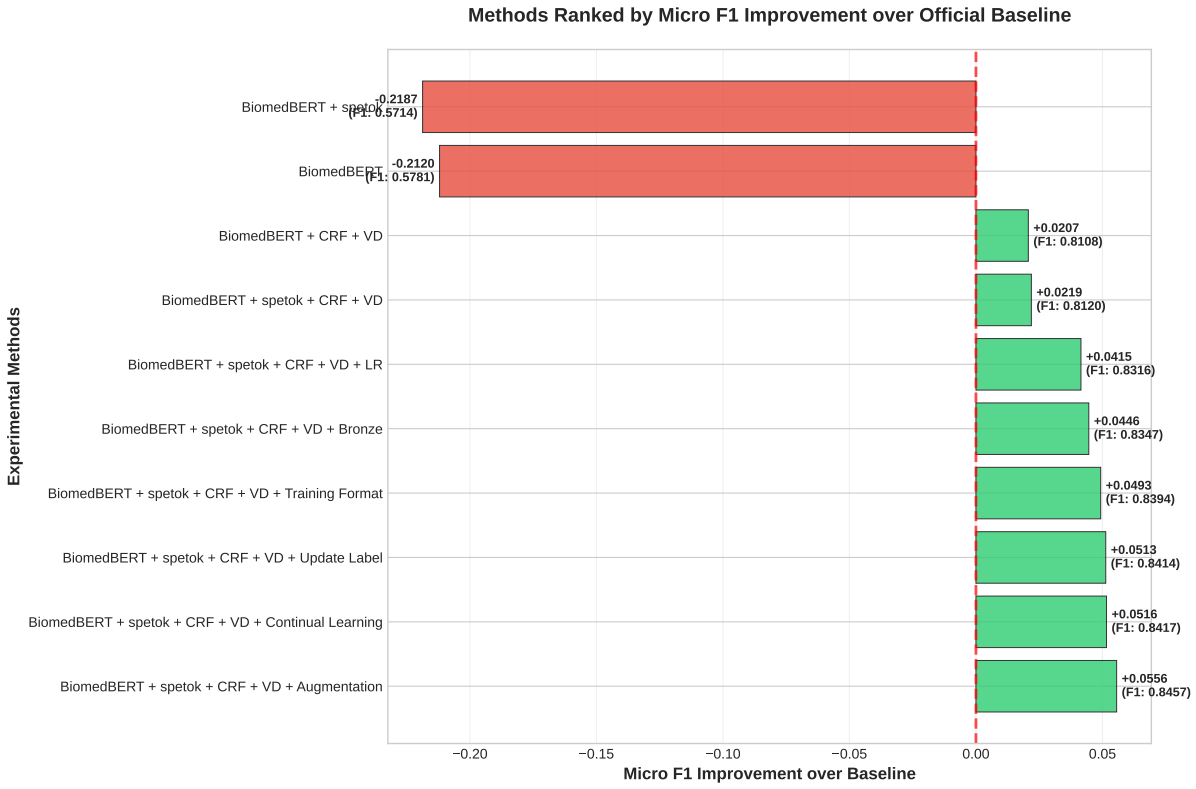


Figure 3: Methods ranking by the percentage of improvement.

1. **CRF**: Adding a CRF layer significantly boosted the result, from Micro F1 at 0.5781 to 0.8028, a total of 22.47% improvement. Learned transition probabilities between adjacent tags efficiently fixed unreasonable span errors and semantic problems stemmed from the token-level predictions of the BiomedBERT model.
2. **Seed Variation Impact**: The choice of random seed significantly affected performance, with seed 17 consistently outperforming compared to seeds 42 and 11. This observation highlights the importance of proper initialization and the need for ensemble approaches to mitigate seed-dependent variance.
3. **Data Format Optimization**: From direct concatenation of abstract and title, to adapting unified text processing `[CLS][title]...[abstract]`, then to separate processing `[CLS][title]...[CLS][abstract]...` resulted in substantial improvements across all metrics. This modification increased macro-F1 scores by approximately 2%.
4. **Bronze Data Integration**: Including Bronze-quality data in training generally improved performance, with consistent gains observed across different seeds (average macro-F1 improvement of $\sim$1%).
5. **Ensemble Effectiveness**: The ensemble approach combining models from three different seeds achieved the highest overall performance, demonstrating the value of model averaging for sequence labeling tasks.
6. **Data Augmentation**: Leveraging additional PubMed abstracts on related topics, pseudo-labeled by our previous best ensemble model, produced our best-performing model as evaluated by Micro-F1, while also achieving the highest Micro-F1, Macro-F1 and Micro-Precision on the leaderboard.

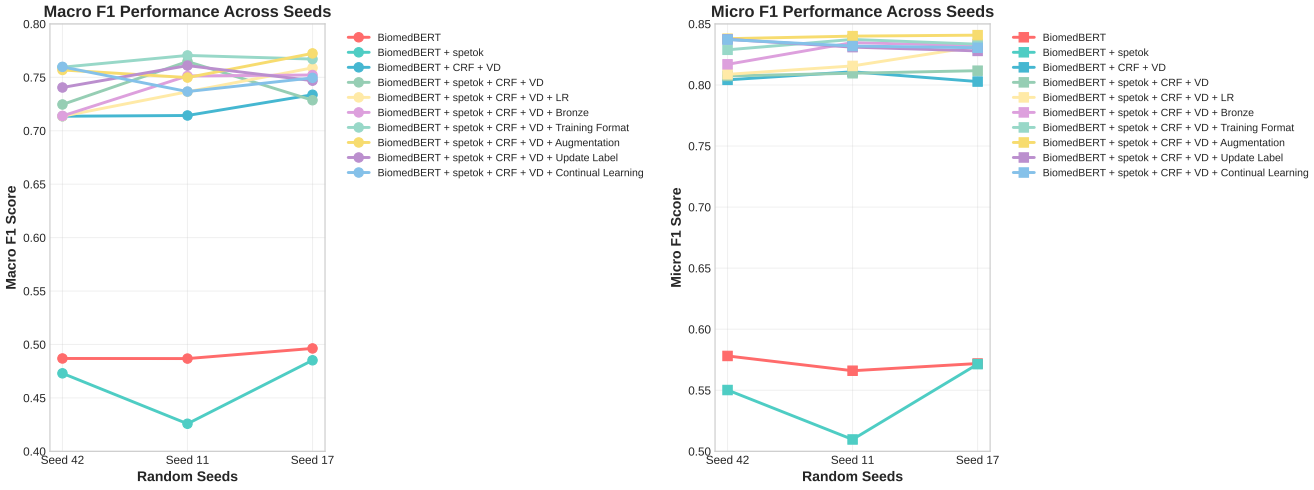


Figure 4: Comparison of Random Seeds

Figure 2 and Figure 3 depicts the progressive improvement of our experimental methods over the official baseline, with the bar charts illustrating the gains in Micro-F1 scores. Figure 4 further shows the performance across different random seeds (42, 11, 17), plotting both Macro F1 and Micro F1 scores for each experimental configuration, and highlighting how seed selection can influence outcomes. Figure 5 provides a comparative view of the ensemble performance, combining models trained with different seeds leads to more robust and superior F1 scores compared to individual models. Our iterative refinements, from strategies like CRF integration, data format optimization, to ensembling has significant impact on the final performance.

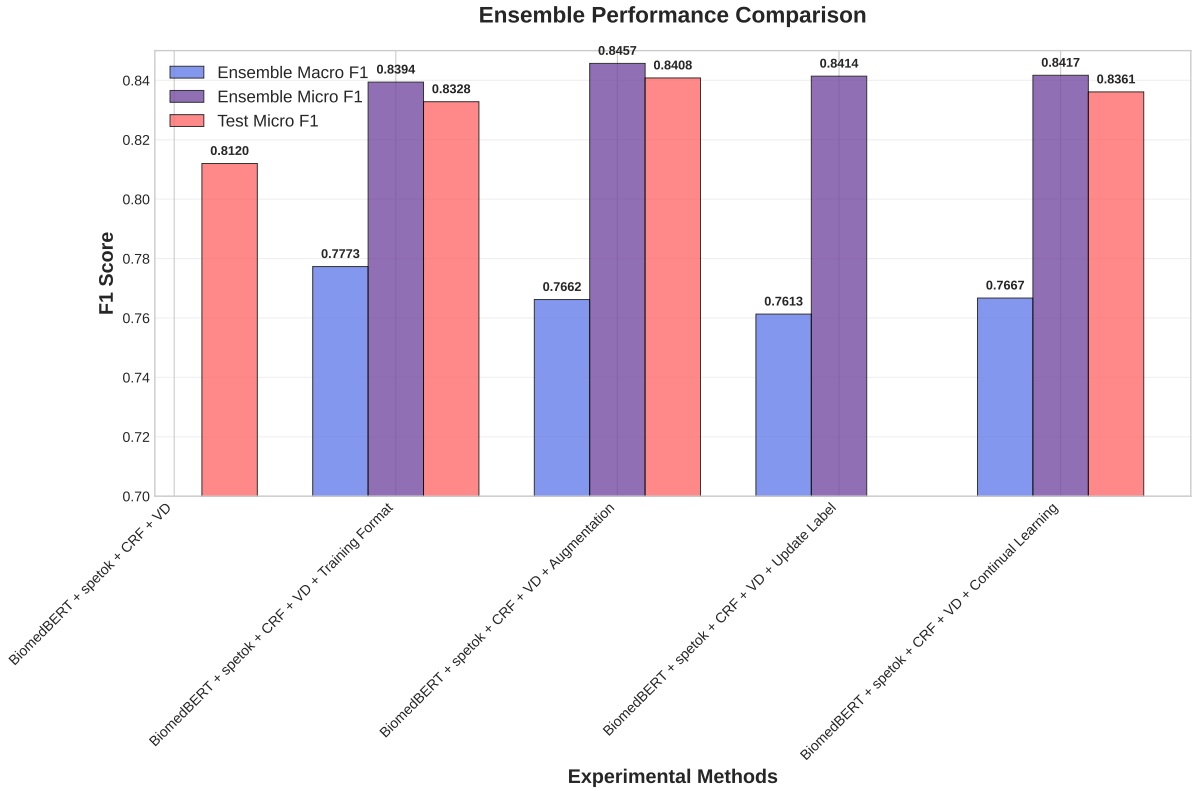


Figure 5: Ensemble Performance

6. Conclusion

The GutUZH team secured the top position among 17 groups, with our best model achieving a Micro-F1 score of 0.8408 on the official test set. This result represents a significant 4.8% improvement over the official Micro-F1 baseline provided by the organizers and a 26.27% absolute improvement over our initial baseline model.

The main achievements and findings of our work include:

- Strong Baseline with Domain-Specific Pre-training
  Our results reaffirm that domain-specific pre-trained models like BiomedBERT serve as powerful foundations for specialized medical NER tasks.
- Impact of Architectural Enhancements
  The integration of a Conditional Random Field (CRF) layer was pivotal, dramatically improving sequence labeling consistency and boosting the Micro-F1 score from 0.5781 to 0.8028. This addition effectively addressed issues of incomplete spans and semantic misclassifications arising from independent token-level predictions.
- Iterative Refinement and Error Analysis
  An error-driven iterative development approach was crucial for optimizing performance. This included meticulous analysis of localization errors, semantic misclassifications, and BIO sequence violations.
- Effective Training Strategies
  The differential learning rates for various model components optimized their convergence. Furthermore, strategic data quality control, including the initial use of high-quality data (Platinum, Gold, Silver) and the later incorporation and iterative improvement of Bronze-level data, proved beneficial.
- Value of Data Augmentation and Ensembling

Data augmentation through pseudo-labeling on additional PubMed articles and ensemble methods, specifically averaging emission and transition matrices from models trained with different random seeds, enhanced robustness and F1 scores.

- Input Representation and Inference Optimization
  Modifying the input format to distinctly mark and process title and abstract segments, along with splitting oversized inputs during inference, contributed to performance gains by providing clearer contextual cues and mitigating information loss from truncation.

For the experiments in the future, while BiomedBERT+CRF proved highly effective, exploring more recent and larger Transformer architectures or even fine-tuning Large Language Models (LLMs) for this biomedical subdomain, despite their current computational costs in high-accuracy settings, could yield further improvements. Experimenting with more diverse ensemble strategies, such as stacking or blending different model architectures or incorporating models trained on different subsets of data, might lead to more robust predictions.

Acknowledgments

Thanks to Dr. Simon Clematide, and Andrianos Michail for giving us invaluable feedbacks throughout the process.

Declaration on Generative AI

During the preparation of this work, the authors used Claude, Gemini, and Deepseek in order to: Improve writing style. After using these tools, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] GutBrainIE Organizers, Gutbrainie @ clef 2025 guidelines - gut brain information extraction - hereditary - unipd, https://hereditary.dei.unipd.it/challenges/gutbrainie/2025/, 2025. Accessed: 2025-05-25.

[2] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, volume TBA of *Lecture Notes in Computer Science*, Springer, 2025, p. TBA.

[3] M. Martinelli, G. Silvello, V. Bonato, G. M. Di Nunzio, N. Ferro, O. Irrera, S. Marchesin, L. Menotti, F. Vezzani, Overview of GutBrainIE@CLEF 2025: Gut-Brain Interplay Information Extraction, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[4] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1–23. doi:10.1145/3458754. arXiv:2007.15779.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, volume 30, 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[6] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2019) 1234 – 1240. doi:10.1093/bioinformatics/btz682.

[7] V. K. Keloth, Y. Hu, Q. Xie, X. Peng, Y. Wang, A. Zheng, M. Selek, K. Raja, C.-H. Wei, Q. Jin, Z. Lu, Q. Chen, H. Xu, Advancing entity recognition in biomedicine via instruction tuning of large language models, Bioinformatics 40 (2024) btae163. URL: https://doi.org/10.1093/bioinformatics/btae163. doi:10.1093/bioinformatics/btae163.

[8] O. Rohanian, M. Nouriborji, S. Kouchaki, F. Nooralahzadeh, L. Clifton, D. A. Clifton, Exploring the effectiveness of instruction tuning in biomedical language processing, Artificial Intelligence in Medicine 158 (2024) 103007. URL: https://doi.org/10.1016/j.artmed.2024.103007. doi:10.1016/j.artmed.2024.103007.

[9] Y. Hu, Q. Chen, J. Du, X. Peng, V. K. Keloth, X. Zuo, Y. Zhou, Z. Li, X. Jiang, Z. Lu, K. Roberts, H. Xu, Improving large language models for clinical named entity recognition via prompt engineering, Journal of the American Medical Informatics Association 31 (2024) 1812–1820. URL: https://doi.org/10.1093/jamia/ocad259. doi:10.1093/jamia/ocad259.

[10] M. S. Obeidat, M. S. A. Nahian, R. Kavuluru, Do llms surpass encoders for biomedical ner?, 2025. URL: https://doi.org/10.48550/arXiv.2504.00664. doi:10.48550/arXiv.2504.00664. arXiv:2504.00664.

[11] C. Peng, X. Yang, K. E. Smith, Z. Yu, A. Chen, J. Bian, Y. Wu, Model tuning or prompt tuning? a study of large language models for clinical concept and relation extraction, Journal of Biomedical Informatics 153 (2024) 104630. URL: https://doi.org/10.1016/j.jbi.2024.104630. doi:10.1016/j.jbi.2024.104630.

[12] K. L. Thant, K. Nongpong, Y. K. Thu, T. Aung, K. H. Wai, T. M. Oo, myner: Contextualized burmese named entity recognition with bidirectional lstm and fasttext embeddings via joint training with pos tagging, 2025. URL: https://doi.org/10.48550/arXiv.2504.04038. doi:10.48550/arXiv.2504.04038. arXiv:2504.04038, to appear in Proceedings of IEEE ICCI-2025.

[13] S. K. Sahu, A. Anand, Unified neural architecture for drug, disease and clinical entity recognition, 2017. URL: https://doi.org/10.48550/arXiv.1708.03447. doi:10.48550/arXiv.1708.03447. arXiv:1708.03447, 23 pages, 2 figures.

[14] M. B. Shishehgarkhaneh, R. C. Moehler, Y. Fang, A. A. Hijazi, H. Aboutorab, Transformer-based named entity recognition in construction supply chain risk management in australia, 2023. URL: https://doi.org/10.48550/arXiv.2311.13755. doi:10.48550/arXiv.2311.13755. arXiv:2311.13755, this work has been submitted to the IEEE for possible publication.

[15] A. E. Mekki, A. E. Mahdaouy, M. Akallouch, I. Berrada, A. Khoumsi, UM6P-CS at SemEval-2022 task 11: Enhancing multilingual and code-mixed complex named entity recognition via pseudo labels using multilingual transformer, 2022. URL: https://doi.org/10.48550/arXiv.2204.13515. doi:10.48550/arXiv.2204.13515. arXiv:2204.13515, submitted to SemEval-2022 Task 11.

[16] F. Pala, M. Y. Akpınar, O. Deniz, G. Eryiğit, Vibertgrid bilstm-crf: Multimodal key information extraction from unstructured financial documents, 2023. URL: https://doi.org/10.48550/arXiv.2409.15004. doi:10.48550/arXiv.2409.15004. arXiv:2409.15004, accepted at MIDAS Workshop, ECML PKDD 2023.

[17] V. Moscato, M. Postiglione, G. Sperlì, A. Vignali, ALDANER: Active Learning based Data Augmentation for Named Entity Recognition, Knowledge-Based Systems 305 (2024) 112682. URL: https://doi.org/10.1016/j.knosys.2024.112682. doi:10.1016/j.knosys.2024.112682.

[18] S. Ghosh, U. Tyagi, M. Suri, S. Kumar, S. Ramaneswaran, D. Manocha, Aclm: A selective-denoising based generative data augmentation approach for low-resource complex ner, ArXiv abs/2306.00928 (2023). doi:10.48550/arXiv.2306.00928.

[19] W. Zhu, J. Liu, J. Xu, Y. Chen, Y. Zhang, Improving low-resource named entity recognition via label-aware data augmentation and curriculum denoising, in: S. Li, R. Wang, K.-Y. Su, D. Ji, Y. Zhang (Eds.), Chinese Computational Linguistics (CCL 2021), volume 12869 of *Lecture Notes in Computer Science*, Springer, Cham, 2021, pp. 292–304. URL: https://doi.org/10.1007/978-3-030-84186-7_24. doi:10.1007/978-3-030-84186-7_24.

[20] A. Romano, G. Riccio, M. Postiglione, V. Moscato, Identifying cardiological disorders in spanish via data augmentation and fine-tuned language models, in: Conference and Labs of the Evaluation Forum, 2024. URL: https://api.semanticscholar.org/CorpusID:271839396.

[21] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. B. A. McDermott, Publicly available clinical bert embeddings, 2019. URL: https://arxiv.org/abs/1904.03323. doi:10.48550/arXiv.1904.03323. arXiv:1904.03323, clinical Natural Language Processing (ClinicalNLP) Workshop at NAACL 2019.

[22] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240. URL: https://arxiv.org/abs/1901.08746. doi:10.1093/bioinformatics/btz682. arXiv:1901.08746.

[23] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019. URL: https://arxiv.org/abs/1910.01108. doi:10.48550/arXiv.1910.01108. arXiv:1910.01108, accepted at the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019.

[24] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. URL: https://arxiv.org/abs/2111.09543. doi:10.48550/arXiv.2111.09543. arXiv:2111.09543, published at ICLR 2023.

[25] M. Moradi, K. Blagec, F. Haberl, M. Samwald, Gpt-3 models are poor few-shot learners in the biomedical domain, ArXiv abs/2109.02555 (2021).

[26] C. Sanchez, Z. Zhang, The effects of in-domain corpus size on pre-training bert, ArXiv abs/2212.07914 (2022). doi:10.48550/arXiv.2212.07914.

[27] J. Yang, X. Hu, W. Huang, H. Yuan, Y. Shen, G. Xiao, Advancing domain adaptation of bert by learning domain term semantics (2023) 12–24. doi:10.1007/978-3-031-40292-0_2.

[28] Z. Li, Q. Wei, L. chin Huang, J. Li, Y. Hu, Y.-S. Chuang, J. He, A. Das, V. Keloth, Y. Yang, C. S. Diala, K. E. Roberts, C. Tao, X. Jiang, W. J. Zheng, H. Xu, Ensemble pretrained language models to extract biomedical knowledge from literature, Journal of the American Medical Informatics Association : JAMIA 31 (2024) 1904 – 1911. doi:10.1093/jamia/ocae061.

[29] J. Li, Q. Wei, O. Ghiasvand, M. Chen, V. Lobanov, C. Weng, H. Xu, Study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora, 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI) (2021) 511–512. doi:10.1109/ICHI52183.2021.00095.

[30] L. Fang, Q. Chen, C.-H. Wei, Z. Lu, K. Wang, Bioformer: an efficient transformer language model for biomedical text mining, ArXiv (2023). doi:10.48550/arXiv.2302.01588.